W254–W257 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl207

SUMOsp: a web server for sumoylation site prediction

Yu Xue[1], Fengfeng Zhou[2], Chuanhai Fu[1,3], Ying Xu[2,*] and Xuebiao Yao[1,3,*]

[1]Laboratory of Cellular Dynamics, Hefei National Laboratory for Physical Sciences, and the University of Science and Technology of China, Hefei, China 230027, [2]Computational Systems Biology Laboratory, Department of Biochemical and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA and [3]Department of Physiology and Cancer Research Program, Morehouse School of Medicine, Atlanta, GA 30310, USA

Received January 27, 2006; Revised March 13, 2006; Accepted March 22, 2006

ABSTRACT

Systematic dissection of the sumoylation proteome is emerging as an appealing but challenging research topic because of the significant roles sumoylation plays in cellular dynamics and plasticity. Although several proteome-scale analyzes have been performed to delineate potential sumoylatable proteins, the bona fide sumoylation sites still remain to be identified. Previously, we carried out a genome-wide analysis of the SUMO substrates in human nucleus using the putative motif ψ-K-X-E and evolutionary conservation. However, a highly specific predictor for *in silico* prediction of sumoylation sites in any individual organism is still urgently needed to guide experimental design. In this work, we present a computational system SUMOsp—SUMOylation Sites Prediction, based on a manually curated dataset, integrating the results of two methods, GPS and MotifX, which were originally designed for phosphorylation site prediction. SUMOsp offers at least as good prediction performance as the only available method, SUMOplot, on a very large test set. We expect that the prediction results of SUMOsp combined with experimental verifications will propel our understanding of sumoylation mechanisms to a new level. SUMOsp has been implemented on a freely accessible web server at: http://bioinformatics.lcd-ustc.org/sumosp/.

INTRODUCTION

Sumoylation, a reversible post-translational modification (PTM) of proteins by the small ubiquitin-related modifiers (SUMOs), is crucial in a variety of biological processes, including transcription (1,2), mRNA metabolism (3), signal transduction (4) and may be involved in the perception of sound (5). Protein sumoylation has also been reported to play essential roles in various diseases and disorders, such as type-1 diabetes (T1D) (6) and Parkinson's disease (PD) (7). SUMO proteins are highly conserved across eukaryotes, and consist of four components in mammals, SUMO-1, SUMO-2, SUMO-3 and SUMO-4 (8). There is only one SUMO gene SMT3 in budding yeast, while there exist at least eight SUMO paralogs in plants (9).

Sumoylation is an unusual phenomenon with quite distinct characteristics. For example, although there are many lysines (K) in a sumoylated protein, only a few of them could be bona fide sumoylation sites. Many sumoylation sites follow a consensus motif ψ-K-X-E (ψ is a hydrophobic amino acid) (8,10) or ψ-K-X-E/D (11,12); however, the accumulating experimental data has shown that about 23% (56/239) of real sumoylation sites don't follow the above consensus motif [Supplementary Table S1 (A)]. It has also been proposed that a nuclear localization signal (NLS) and a consensus motif confer the ability to be sumoylated. But there exist some real SUMO substrates that are not localized in nucleus. For example, protein DRP1 (dynamin related protein) is localized in the mitochondria and is sumoylated during mitochondrial fission (13). In this regard, our understanding of sumoylation mechanisms is still in its infancy. Moreover, the sumoylation process is dynamic and only a small fraction of the proteome, often <1%, will be sumoylated *in vivo* at any given time (10).

These complex features of sumoylation sites have introduced great difficulties in the systematic analysis of the sumoylation proteome. Using mass spectrometry (MS) approaches, several large-scale experiments of sumoylation substrates have been carried out (12,14–17), however, the *bona fide* sumoylation sites still remain to be identified. In this regard, computational approaches might represent a promising method for identification of sumoylation sites.

*To whom correspondence should be addressed. Tel: +86 551 3606304; Fax: 001 404 752 1045; Email: yaoxb@ustc.edu.cn
*Correspondence may also be addressed to Ying Xu. Tel: +001 706 542 9779; Fax: +001 706 542 9751; Email: xyn@bmb.uga.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Previous work on *in silico* identification of SUMO substrates with their sumoylation sites is mainly based on identification of the consensus motif, ψ-K-X-E or ψ-K-X-E/D, which may miss many true positives. And since many consensus sites are not sumoylated, these approaches will often generate very high false positive prediction rates. In this work, we have developed a computational system, SUMOsp—SUMOylation Sites Prediction, based on two methods, GPS (18,19) and MotifX (20). GPS and MotifX are originally designed for phosphorylation site prediction, and leave-one-out validation and 5-fold cross validation in this article indicate that these two pattern recognition strategies are also robust and accurate for the sumoylation site prediction. SUMOsp offers at least as good prediction performance as the only existing system, SUMOplot. To facilitate applications of this system by other users, we have developed an easy-to-use web server of SUMOsp, which is freely accessible at: http://bioinformatics.lcd-ustc.org/sumosp/.

IMPLEMENTATION

Data preparation

We searched PubMed with keywords 'SUMO' and 'sumoylation', and manually curated 239 unambiguously experimentally-identified sumoylation sites in 144 proteins from ~400 research articles published online before December 10, 2005. We have retrieved their primary sequences from Swiss-Prot/TrEMBL database (http://cn.expasy.org). Due to the database updates, the sumoylation positions reported in the literature may have changed in the current primary sequences, therefore the dataset was manually validated before our analyzes.

Algorithm

We first define a potential sumoylation peptide *PSP*(*n*) as a lysine (K) residue flanked by *n* residues upstream and *n* residues downstream. We hypothesize that the biochemical properties of a sumoylation site mainly depend on the neighboring amino acids, and this hypothesis has been satisfactorily confirmed by our validation results. In this work, we use $n = 7$ for *PSP*(*n*)'s, which is confirmed by the prediction performance to be sufficient to represent the flanking information of a sumoylation site. Although other matrices could be employed, we choose BLOSUM62 as we have previously used (19).

In this study, we have employed two powerful prediction strategies, GPS (18,19) and MotifX (20), for prediction of sumoylation sites, and our server provides both results to its users.

As described in (19), two peptides flanking the same amino acid may have similar PTM, if the BLOSUM62 substitution score between them is sufficiently high. In this study, GPS firstly partitioned the dataset of *PSP*(7) flanking the 239 known sumoylation sites into three clusters. For a given *PSP*(7) flanking a lysine (K) amino acid and one of the clusters, the averaged value of the scores between this peptide and the peptides in the cluster is defined as the score of this cluster. The GPS score of this given peptide is defined as the maximum one of the scores between the peptide and the clusters. We use a particular cut-off value to make the final judgment.

MotifX (20) generated a set of highly-specific motifs for the sumoylation sites, IKXEP, VKXE, IKXE, LKXE and KXE (X can be any amino acid), which can be easily used by users. In fact, we found that MotifX exhibits greater computing power when it combines with GPS. For example, a combination of MotifX with GPS predicts *PSP*(7) as a positive hit when the peptide is predicted as positive for either of them. So SUMOsp, the integration of GPS and MotifX, acts in this way.

RESULTS

We use sensitivity (Sn), specificity (Sp) and accuracy (Ac) to evaluate the performance of SUMOsp. Sensitivity and specificity measure the positive and negative predictions, respectively, while accuracy provides the correct prediction ratio. It is worth noting that we found that these measures are inadequate for the cases where the numbers of positive and negative data differ significantly. So in addition to Sn, Sp and Ac values, we have also used a correlation coefficient (CC) to assess our prediction system. CC is between −1 and 1, and the larger a CC is, the more accurate the prediction is.

Analogous to the previous work (18,19,21), the known sumoylation sites are regarded as the positive data, while all the other lysine (K) amino acids in the known sumoylation substrates are regarded as the negative data. Among the data with positive predictions by SUMOsp, the real positive ones are called true positives (TP), and the others are called false positives (FP). Among the data with negative predictions by SUMOsp, the real positive ones are called false negatives (FN), while the others are called true negatives (TN).

The performance measurements sensitivity (Sn), specificity (Sp), accuracy (Ac) and Matthews' correlated coefficient (CC) (22) are defined as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \quad \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}},$$

$$\mathrm{Ac} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}},$$

and

$$\mathrm{CC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}.$$

We provide three cut-off scores, 1.5, 4 and 18, which are only effective for the GPS scores. Users may choose different cut-off score according to their requirements on the prediction performance (refer to Supplementary Table S2). SUMOsp

Table 1. Prediction performance of SUMOsp and SUMOplot

Predictor	Threshold	Ac (%)	Sn (%)	Sp (%)	CC
SUMOsp	18	92.71	83.68	93.08	0.5012
	4	80.43	89.12	80.07	0.3232
SUMOplot	high	89.94	79.50	93.31	0.4825
	all	80.45	88.70	80.07	0.3211

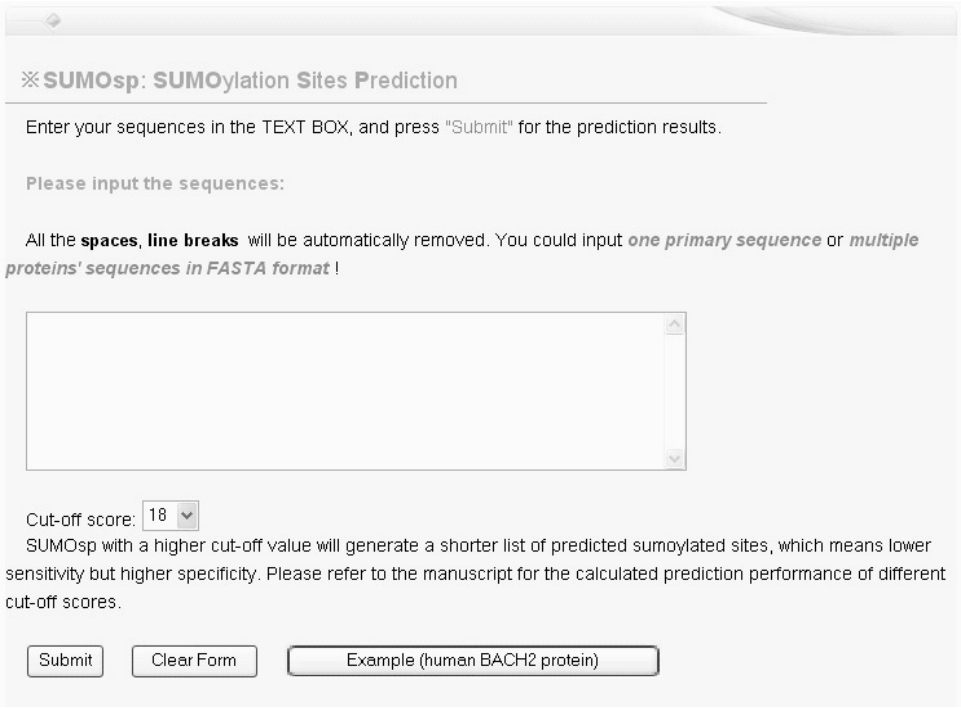

Figure 1. The prediction page of SUMOsp web server.

with cut-off score 0 will generate the prediction results of GPS and MotifX for all the lysines, which is of interest for further investigations.

We have compared the prediction performance of SUMOsp to the only publicly available tool SUMOplot (http://www.abgent.com/doc/sumoplot). Making predictions based on hydrophobic similarity with the consensus motif and the degree of matching with the sumoylation sites from Ubc9-binding substrates, SUMOplot is considered as an excellent computational program. Here we denote the two levels of stringencies of SUMOplot as high (hits with high probability) and all (all predictions). As in Table 1, the Ac, Sn, Sp and CC of SUMOsp with threshold 18 are 92.71%, 83.68%, 93.08% and 0.5012, respectively, while the Ac, Sn, Sp and CC of SUMOsp with threshold 4 are 80.43%, 89.12%, 80.07% and 0.3232, respectively. The Ac, Sn, Sp and CC of SUMOplot at high/all levels are 89.94%/80.45%, 79.50%/88.70%, 93.31%/80.07% and 0.4825/0.3211, respectively. So SUMOsp is more accurate by all measurements. To test SUMOsp's robustness, we have used both Leave-one-out validation and 5-fold cross validation. Both methods show similar levels of performance to the above results. The Ac, Sn, Sp and CC of the consensus motif ψ-K-X-E are 97.21%, 74.48%, 98.16% and 0.6689 respectively. So SUMOsp provides better sensitivity while keeping similar specificity. Experimentalists may want to generate a more reliable *in silico* prediction results by integrating the above methods, phylogenetic conservation and structural analysis. Detailed information about the validations could be found in Supplementary Table S2.

To illustrate how robust SUMOsp is in regard of threshold-independent performance, we provided the receiver operating characteristic (ROC) curves of self validation, Leave-one-out validation and 5-fold cross validation (refer to Supplementary Figure S1). Both the ROC curves and the areas under the ROC curves (AUC) suggest that SUMOsp is a robust prediction system.

For those non-canonical real sumoylation sites, SUMOsp can also provide a satisfying prediction performance [as in Supplementary Table S1 (B)].

USE OF SUMOSP WEB SERVICE

SUMOsp web server has been developed in an easy-to-use manner. A user can visit SUMOsp at http://bioinformatics.lcd-ustc.org/sumosp/prediction.php (Figure 1), enter the protein sequences either in raw format or FASTA format into the text box, and run the program by pressing the 'Submit' button. The prediction results should be regarded as potential sites before experimental validation. And by pressing the word here in the sentence 'Download the TAB-deliminated data file from here', a user can get prediction results in tab-deliminated plain text to be used for further consideration.

DISCUSSION AND CONCLUSION

The systematic identification of the sumoylation proteome represents a great challenge. Although experimental verifications are essential, computational methods can serve as a complementary and powerful tool to help accelerate the sumoylation research. Previously, we have performed a genome-wide analysis of the SUMO substrates in human nucleus, based on pattern recognition and evolutionary conservation (5). An *in silico* predictor for sumoylation sites is still urgently needed.

In this work, we have developed a novel computational method and computer program, SUMOsp, for the

highly-specific prediction of sumoylation sites. Based on its prediction performance, we believe that SUMOsp could serve as a powerful and complementary tool for *in vivo* or *in vitro* sumoylation site identification; and the combination of computational analyzes with experimental verification could greatly speed up our understanding of the mechanisms and dynamics of sumoylation systematically.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Hongmei Wang, Dr Changjiang Jin and Han Yan for insightful discussion during the course of this study. The work is supported by Chinese Natural Science Foundation (39925018, 30270654 and 30270293), Chinese Academy of Science (KSCX2-2-01), Chinese 973 project (2002CB7-13700), Chinese Minister of Education (20020358051), American Cancer Society (RPG-99-173-01) and National Institutes of Health (DK56292; CA92080). X.Y. is a Georgia Cancer Coalition Eminent Scholar. F.Z. and Y.X. work is supported by the Georgia Cancer Coalition, National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204), and Department of Energy's Genomes to Life Program (http://doegenomestolife.org/) under project, 'Carbon Sequestration in *Synechococcus* sp.: From Molecular Machines to Hierarchical Modeling. Special thanks go to the two anonymous reviewers, whose suggestions greatly improved the presentations of our manuscript. Funding to pay the Open Access publication charges for this article was provided by NIH DK56292.

Conflict of interest statement. None declared.

REFERENCES

1. Gregoire,S. and Yang,X.J. (2005) Association with class IIa histone deacetylases upregulates the sumoylation of MEF2 transcription factors. *Mol. Cell Biol.*, **25**, 2273–2287.
2. Girdwood,D.W., Tatham,M.H. and Hay,R.T. (2004) SUMO and transcriptional regulation. *Semin. Cell Dev. Biol.*, **15**, 201–210.
3. Li,T., Evdokimov,E., Shen,R.F., Chao,C.C., Tekle,E., Wang,T., Stadtman,E.R., Yang,D.C. and Chock,P.B. (2004) Sumoylation of heterogeneous nuclear ribonucleoproteins, zinc finger proteins, and nuclear pore complex proteins: a proteomic analysis. *Proc. Natl Acad. Sci. USA*, **101**, 8551–8556.
4. Liang,M., Melchior,F., Feng,X.H. and Lin,X. (2004) Regulation of Smad4 sumoylation and transforming growth factor-beta signaling by protein inhibitor of activated STAT1. *J. Biol. Chem.*, **279**, 22857–22865.
5. Zhou,F., Xue,Y., Lu,H., Chen,G. and Yao,X. (2005) A genome-wide analysis of sumoylation-related biological processes and functions in human nucleus. *FEBS Lett*, **579**, 3369–3375.
6. Li,M., Guo,D., Isales,C.M., Eizirik,D.L., Atkinson,M., She,J.X. and Wang,C.Y. (2005) SUMO wrestling with type 1 diabetes. *J. Mol. Med.*, **83**, 504–513.
7. Shinbo,Y., Niki,T., Taira,T., Ooe,H., Takahashi-Niki,K., Maita,C., Seino,C., Iguchi-Ariga,S.M. and Ariga,H. (2005) Proper SUMO-1 conjugation is essential to DJ-1 to exert its full activities. *Cell Death Differ*, **13**, 96–108.
8. Hay,R.T. (2005) SUMO: a history of modification. *Mol. Cell*, **18**, 1–12.
9. Kurepa,J., Walker,J.M., Smalle,J., Gosink,M.M., Davis,S.J., Durham,T.L., Sung,D.Y. and Vierstra,R.D. (2003) The small ubiquitin-like modifier (SUMO) protein modification system in *Arabidopsis*. Accumulation of SUMO1 and -2 conjugates is increased by stress. *J. Biol. Chem.*, **278**, 6862–6872.
10. Johnson,E.S. (2004) Protein modification by SUMO. *Annu. Rev. Biochem.*, **73**, 355–382.
11. Melchior,F., Schergaut,M. and Pichler,A. (2003) SUMO: ligases, isopeptidases and nuclear pores. *Trends Biochem. Sci.*, **28**, 612–618.
12. Denison,C., Rudner,A.D., Gerber,S.A., Bakalarski,C.E., Moazed,D. and Gygi,S.P. (2004) A proteomic strategy for gaining insights into protein sumoylation in yeast. *Mol Cell Proteomics*, **4**, 246–254.
13. Harder,Z., Zunino,R. and McBride,H. (2004) Sumo1 conjugates mitochondrial substrates and participates in mitochondrial fission. *Curr. Biol.*, **14**, 340–345.
14. Gocke,C.B., Yu,H. and Kang,J. (2005) Systematic identification and analysis of mammalian small ubiquitin-like modifier substrates. *J. Biol. Chem.*, **280**, 5004–5012.
15. Hannich,J.T., Lewis,A., Kroetz,M.B., Li,S.J., Heide,H., Emili,A. and Hochstrasser,M. (2005) Defining the SUMO-modified proteome by multiple approaches in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **280**, 4102–4110.
16. Rosas-Acosta,G., Russell,W.K., Deyrieux,A., Russell,D.H. and Wilson,V.G. (2005) A universal strategy for proteomic studies of SUMO and other ubiquitin-like modifiers. *Mol. Cell Proteomics*, **4**, 56–72.
17. Wykoff,D.D. and O'Shea,E.K. (2005) Identification of sumoylated proteins by systematic immunoprecipitation of the budding yeast proteome. *Mol. Cell Proteomics*, **4**, 73–83.
18. Xue,Y., Zhou,F., Zhu,M., Ahmed,K., Chen,G. and Yao,X. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res.*, **33**, W184–W187.
19. Zhou,F.F., Xue,Y., Chen,G.L. and Yao,X. (2004) GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem. Biophys. Res. Commun.*, **325**, 1443–1448.
20. Schwartz,D. and Gygi,S.P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale datasets. *Nat. Biotechnol.*, **23**, 1391–1398.
21. Kim,J.H., Lee,J., Oh,B., Kimm,K. and Koh,I. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.
22. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.